جامعة مصر للمعلوماتية
EGYPT UNIVERSITY
OF INFORMATICS

Egypt University of Informatics
Computer and Information Systems
Data Analysis Course

Analysing High School Students Grades And Potential Factors

Submitted by: Mahmoud Barbary (21-101128) - Ayman Talaat (21-101129)

Under the Supervision of: Dr. Mohamed Taher Alrefaie

5/23/2023

Introduction

The Thanawaya Amma exam is a high-stakes test that is taken by Egyptian high school students at the end of their final year. The exam is used to determine which universities students are eligible to attend and what majors they can pursue.In this context, it's important to understand what affects a student's grades and a factor like which regions are best-or-worst-performing. Such knowledge can be crucial in identifying the critical points where the education system can be improved upon, and where attention should go to reinforce and support certain regions accordingly.
Thus our aim is to expose underlying unique relations between the different factors that can go into a student's grade, like their city of residence, branch and gender.

Research Question

Is there any difference between the grades in the governorate of Red Sea and the rest of the population?

Hypothesis

The governorate of Red Sea`s mean percentage is different from that of the population.

Population of Interest:

All of Egypt's high school final year students.

Sampling Method:

Our sample is all of Egypt's high school final year students, and their grades for the year 2022. This method is convenience sampling because this dataset is the most recent available one.

Bias Identification:

If we were to make any inferences from this data set, since it is limited by a single year, it would not take into consideration any occurring changes through the years like a change in the education system, the exam form and/or the grading scheme.

Survey Questions/Collected Data/Dataset:

We utilised a Kaggle dataset that includes each subject grades, as well as information on the student like their branch (literature, maths and science), governorate and city of residence, id and gender. This dataset served as the basis for our analysis.

Each row represents a singular student, their id, their grade in each subject and finally whether they have passed , failed or will need to retake a subject or more. Please note that students would have empty values in some subjects depending on their branch.

Number of samples used: 682,349

Kaggle Dataset Link: [High School (ثانوية عامة) Public Results 2022 EG | Kaggle]

Improvements

- In our data visualisation approach, we have made significant improvements to enhance readability and interpretation. Previously, bar plots representing each governorate's student grades were displayed as a lengthy sequence of individual images, which could be overwhelming and less intuitive to navigate. We've now enhanced this by compiling these bar plots into a single, easy-to-view animated GIF. This not only consolidates the visual data representation but also enhances the overall aesthetic appeal of our analysis.

- We tried to make a quick test to see the number of students in each government and how this affects the average grade achieved, in order to test that having a large number of students while have an impact on the grades, as what comes to mind that with an increased quantity of various grades while immensely impact the average grade, while having a smaller amount while result in relatively stable results when calculating the mean.

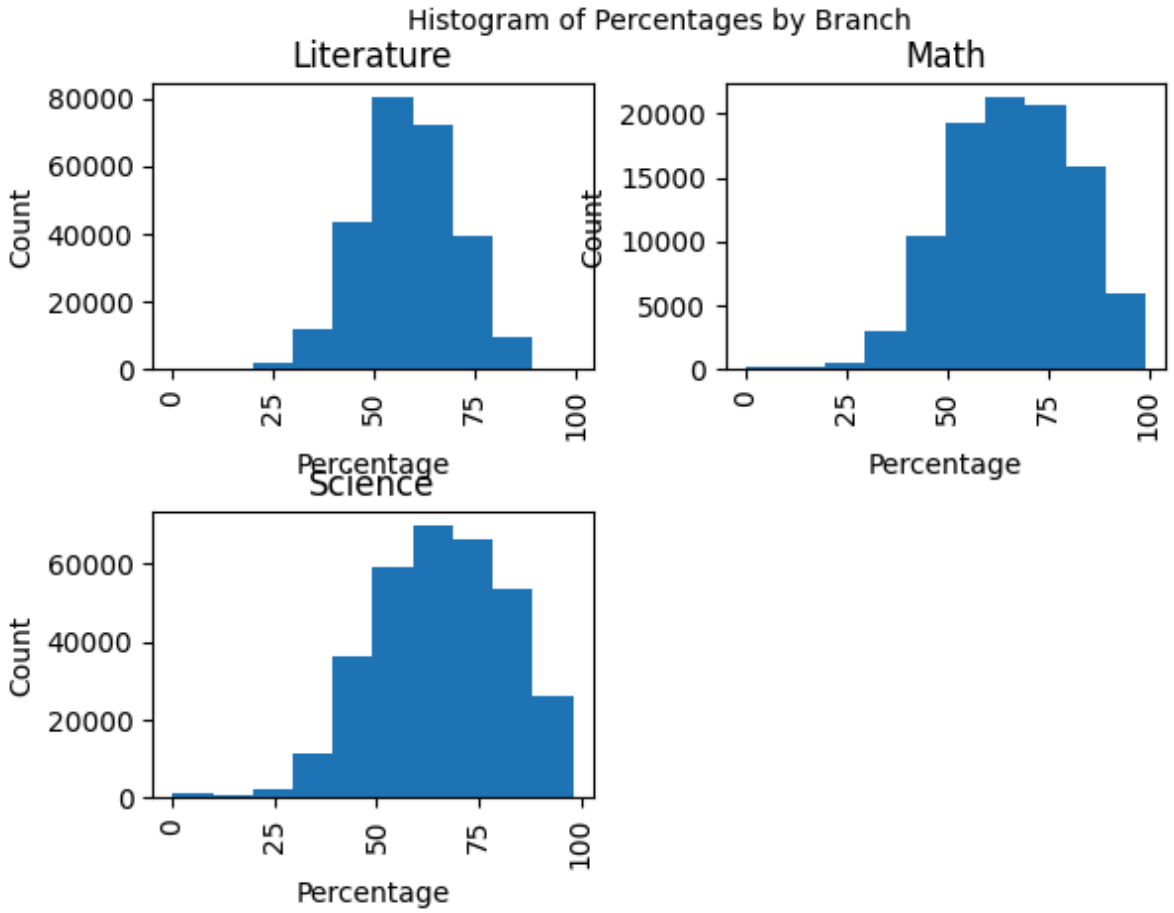
- We've made significant improvements in our geographical data by boosting its visualisation and accuracy levels. In the past, we presented an aggregate view of grade distributions for every subject across all Egyptian governorates on one map. However, such an approach had limitations as it risked masking specific performance patterns within diverse academic sections. To overcome this shortfall, we have now introduced separate maps for each subject area. For example, we've created an exclusive map that focuses on the 'Literature' section comprising subjects like Psychology, Philosophy, History and Geography; allowing us to disentangle individual performances within those areas clearly. Moreover; we made a graph comparing the average grades in each distinct governorate in subjects that are common between all academic sections, which are the language related subjects, like Arabic and English. In addition to a graph that highlights the comparison between the average grades of each governorate in physics and chemistry which are common between maths and science branches under the title of General science. The objective here is to enable a more focused examination and simpler comprehension of the regional grading patterns for every academic segment throughout the land.

Analysis:

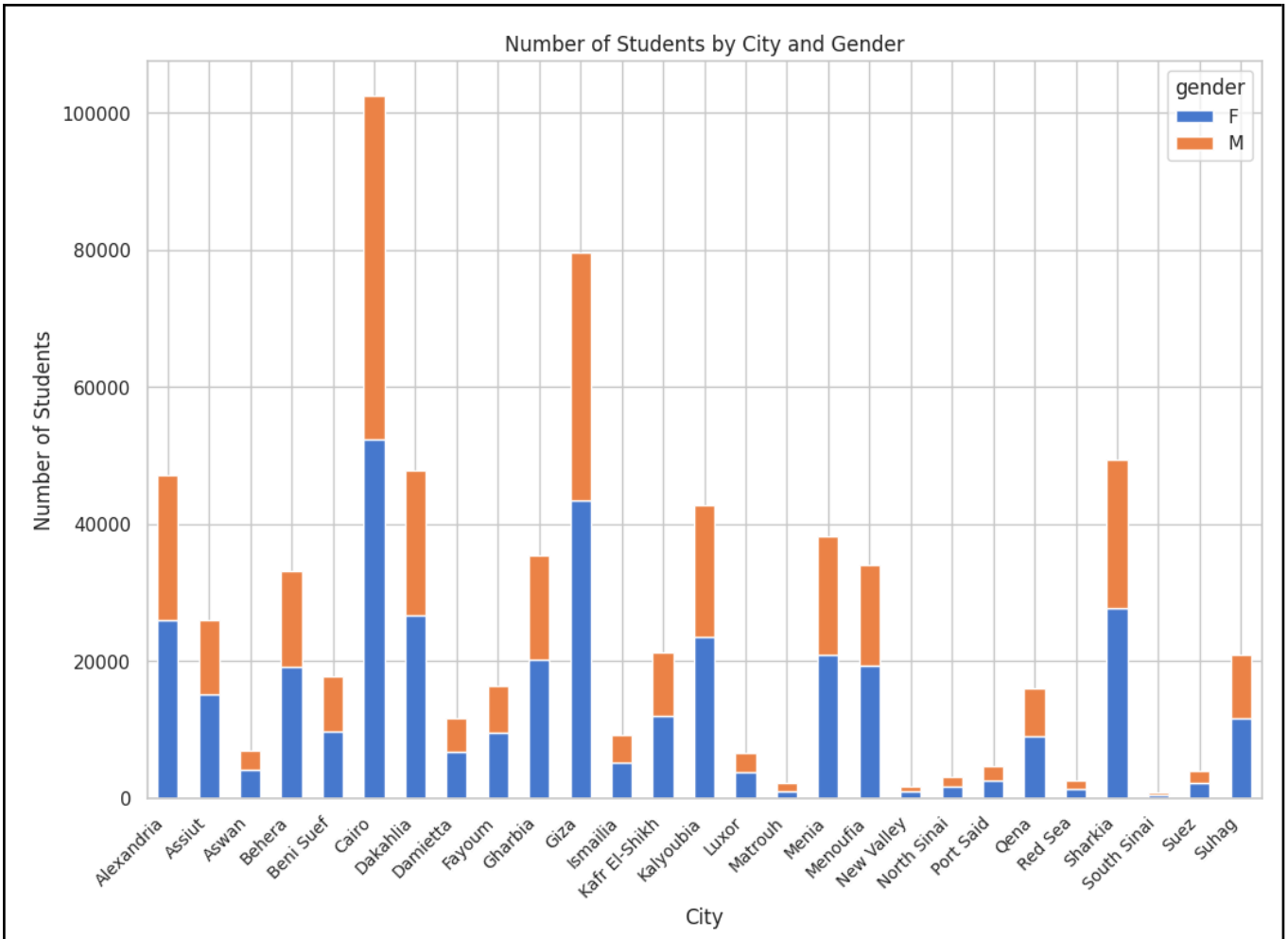
Grade Means			
Arabic	50.06	Psychology	37.72
1st Language	32.91	Chemistry	38.69
Pure Maths	37.21	Biology	35.98
History	32.77	Geology	43.47
Geography	33.13	Applied Maths	39.45
Philosophy	38.10	Physics	35.59
-	Total % Mean	63.14	-

Gender Proportion	
Males	55.4%
Females	44.6%

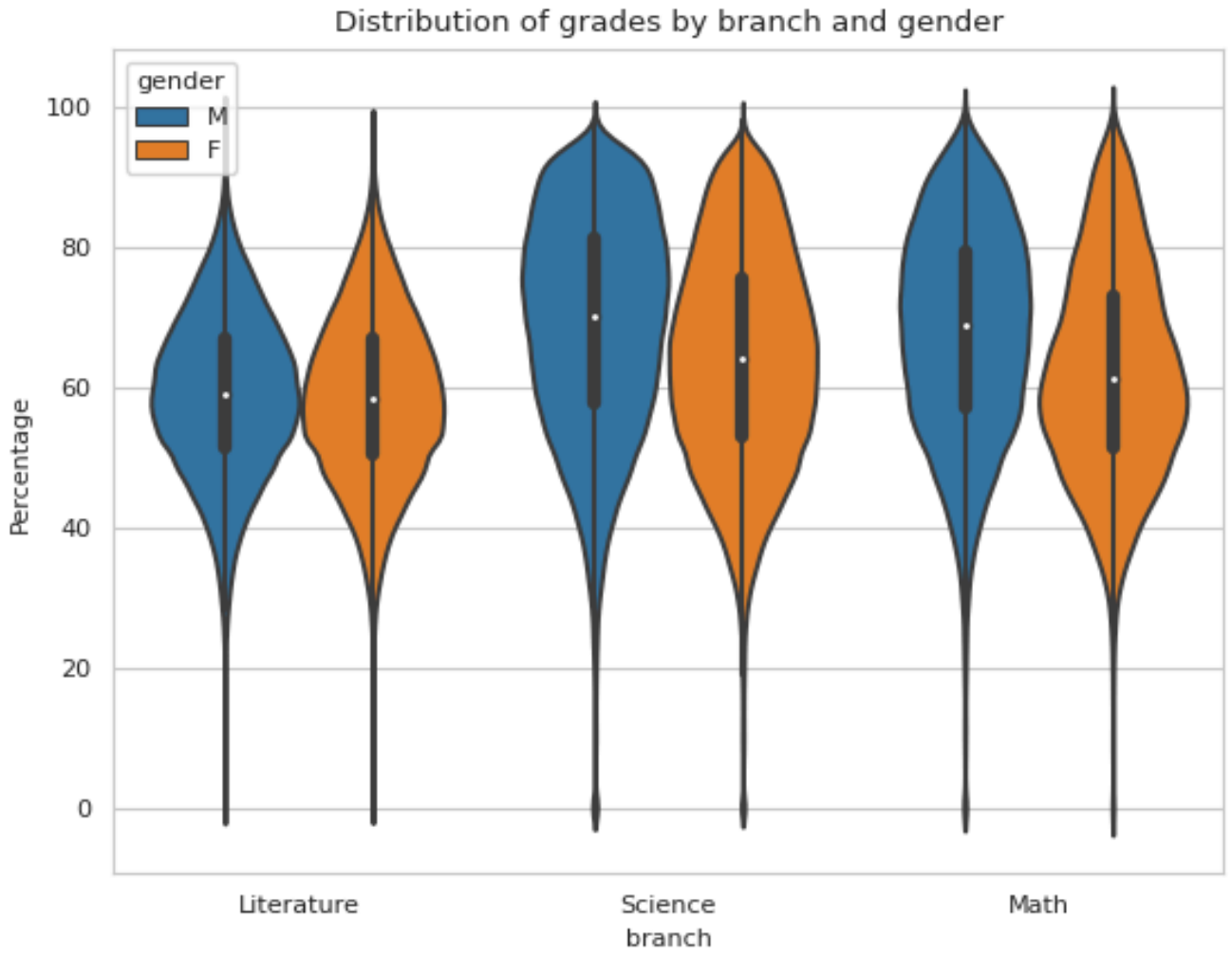
Graphs



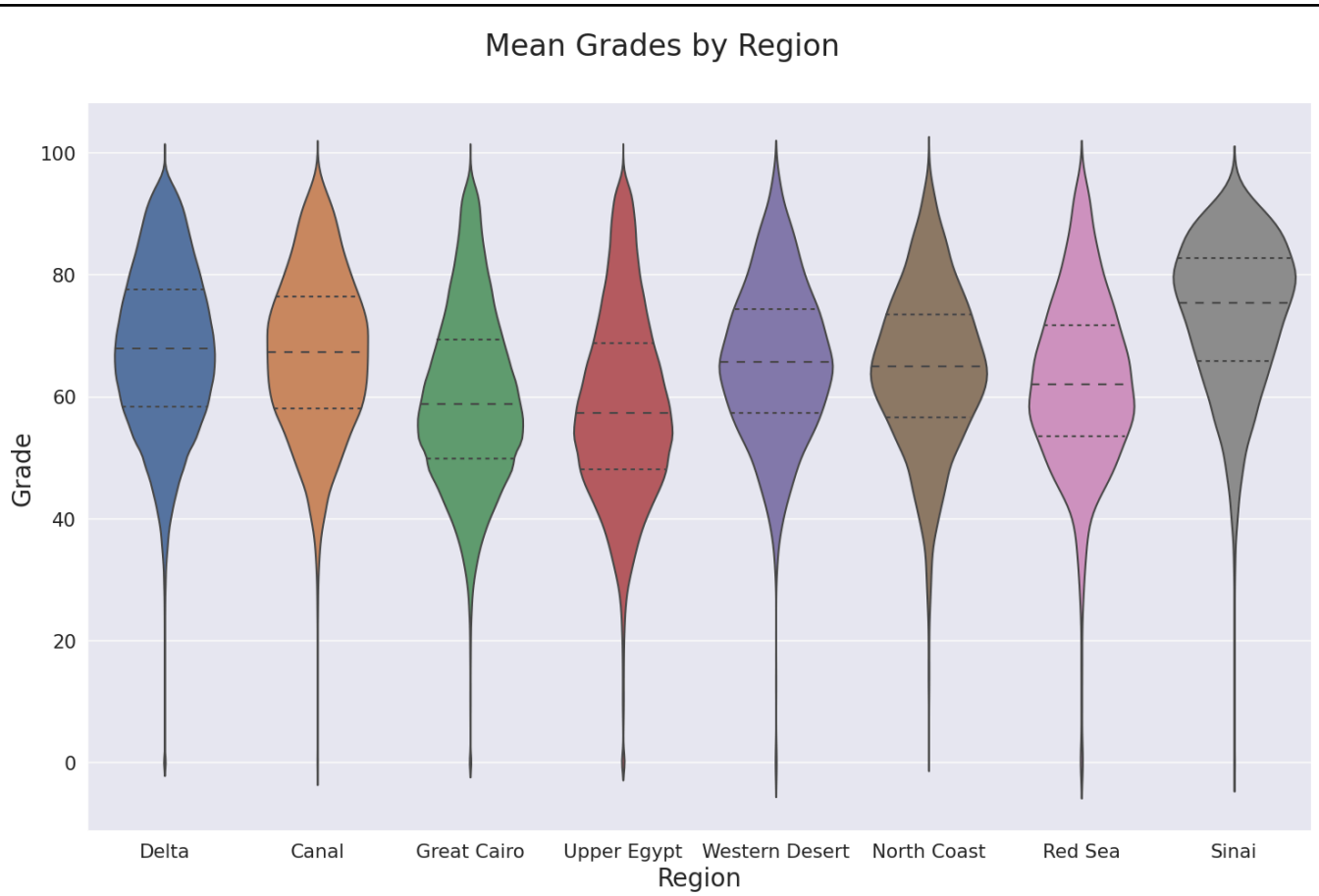
These histograms illustrate the distributions of grades separated by each branch.
It's clear that all of them are kind of close to the shape of a normal distribution, with maths and science branches being quite similar.



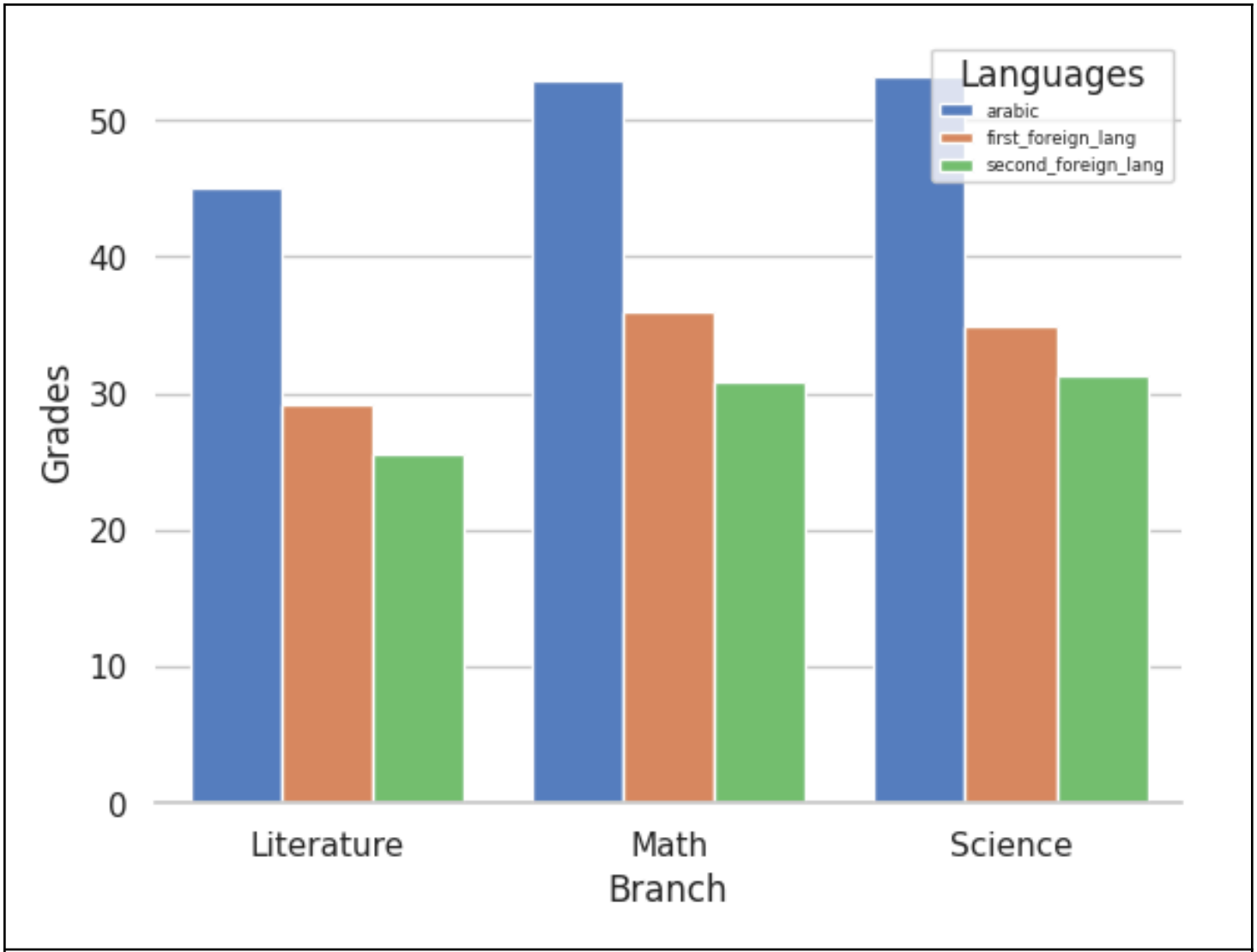
This is a bar chart that shows how many male & female students attended Thanaweya Amma During that year in each city.



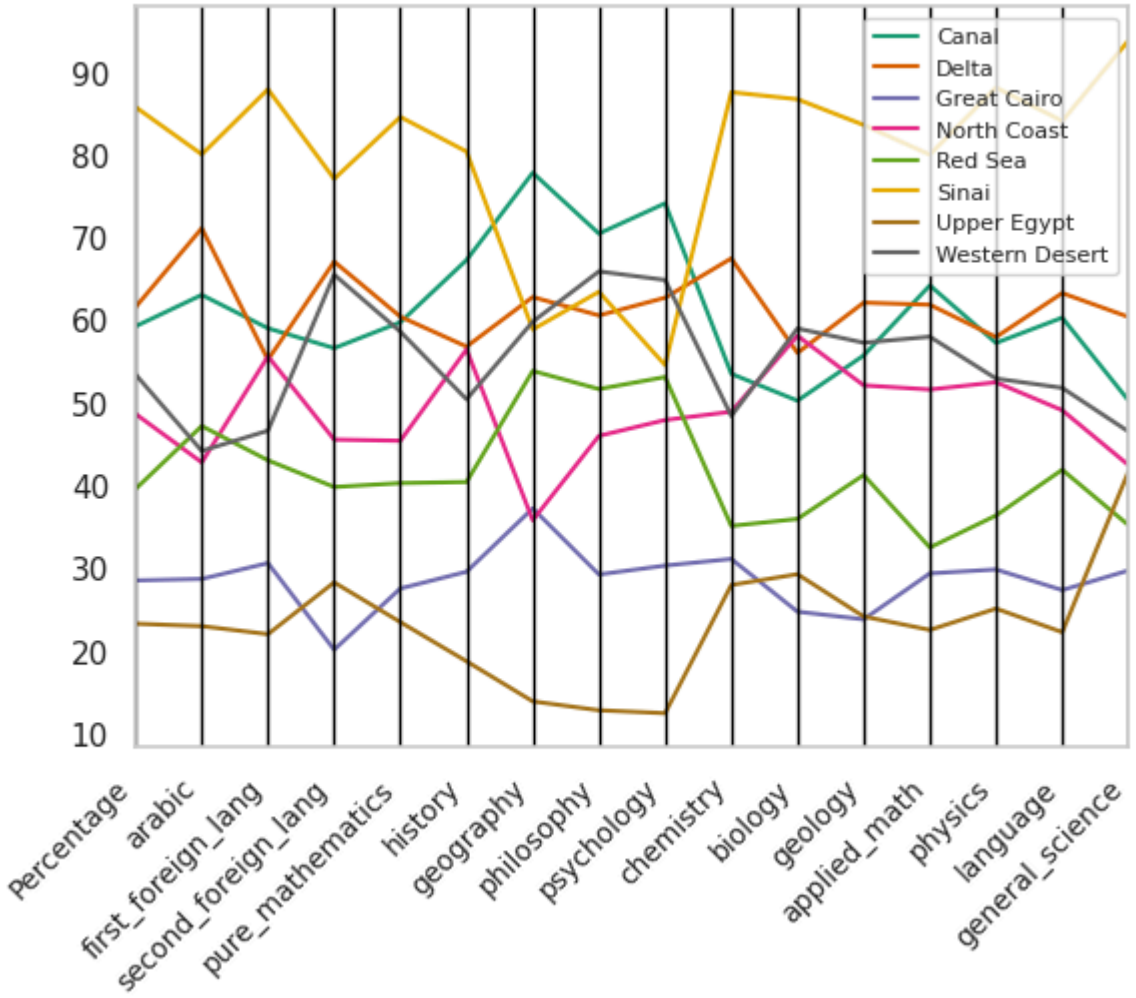
Violin plot showing the difference in grades between males and females in each branch.
In literature, we can see that the shapes are almost identical.
But in science, males seem to perform slightly better on average.
And in mathematics, as seen in the figure, the corresponding plot is comparatively spread out towards lower grades than the male counterpart.



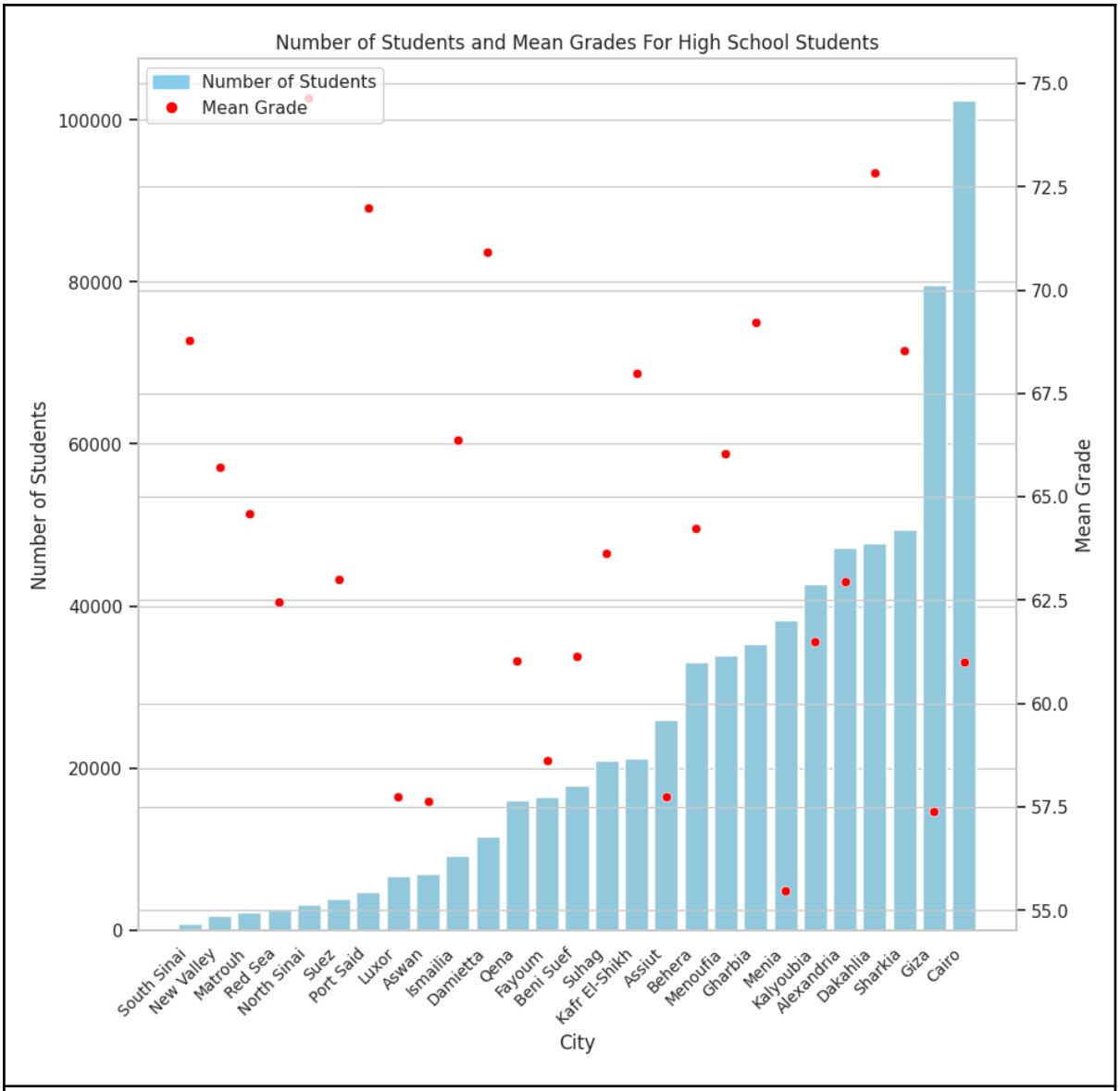
Violin plot showing the distribution of grades in different regions. The plot shows that Great Cairo, Upper Egypt, and Red Sea have the lowest grades respectively. While regions such as Delta ,Canal and North Coast have relatively higher grades. The highest grades are in the Sinai region.



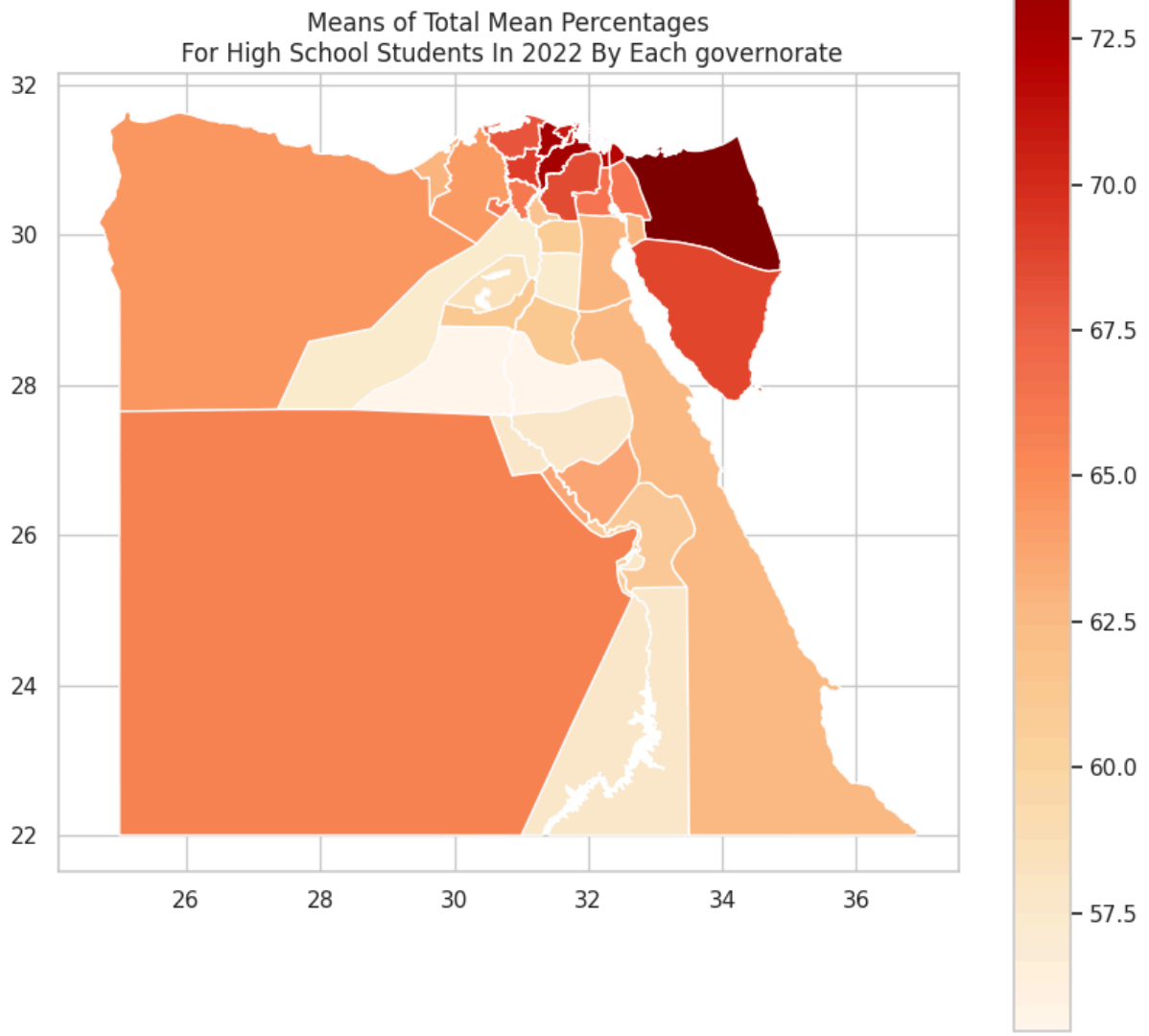
These grouped bar charts show the difference in mean grades in each of the Language Subjects. Contrary to popular belief, literature students who should theoretically perform the best in these subjects as it has a lot more to do with their branch, in reality have the lowest grades in all 3 languages.
Then, as it is common throughout the data, maths and sciences students are nearly identical.



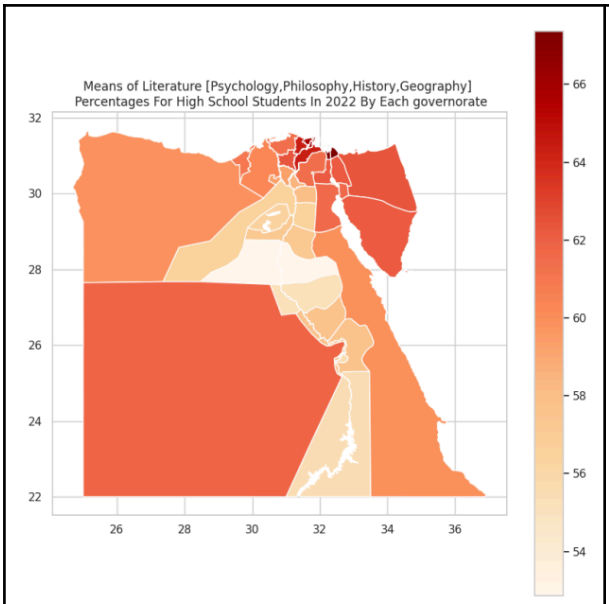
This line plot shows how the mean percentages of students in different regions vary with different subjects. Our most notable observations in this graph are that the Sinai region overtakes the rest of the regions in most subjects while on the other hand of the spectrum Upper Egypt is falling behind.



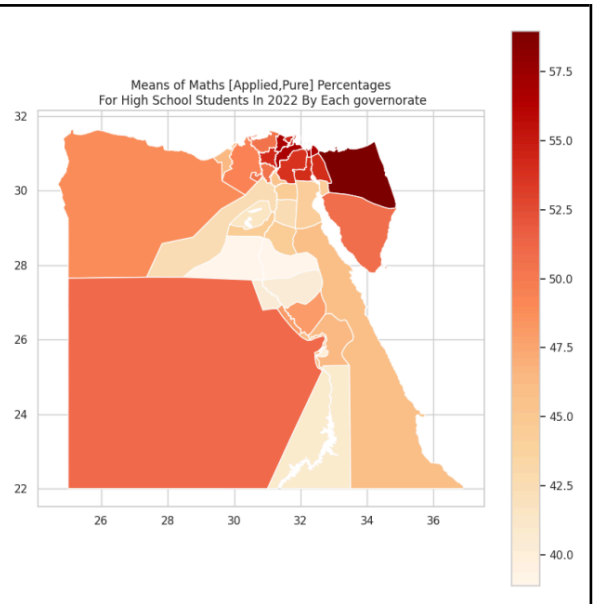
This is a combination of a bar chart and a scatterplot, the bars show how many students took Thanawiya Aama in that city while the dots show the mean grades of all students in each city across Egypt.



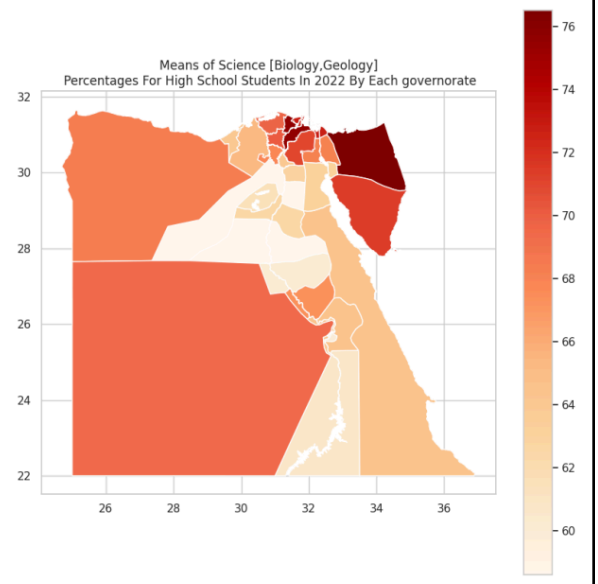
Geographical heat map of Egypt showing the means of percentage grades for each governorate. The warmer the colour, the higher the mean it has. The map shows that North Sinai is the top performer while Minya is unfortunately quite the opposite.
We can see certain clustered regions of multiple governorates have a collective lower average of grades than others and vice versa.



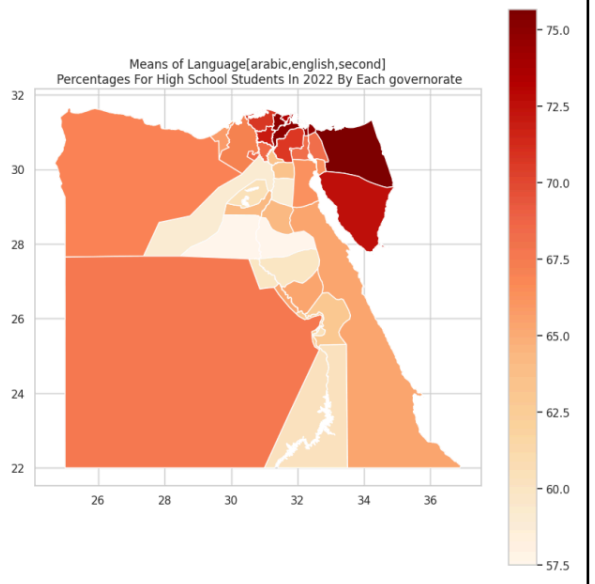
Geographical heat map of Egypt showing the means of Literature students performance in their specified subjects for each governorate. The warmer the colour, the higher the mean it has. The map shows that Port Said is the top performer while Al Minia is unfortunately quite the opposite.
We can see certain clustered regions of multiple governorates have a collective lower average of grades than others and vice versa.



Geographical heat map of Egypt showing the means of Maths students performance in their specified subjects for each governorate. The warmer the colour, the higher the mean it has. The map shows that Northern Sinai is the top performer while Al Minia is unfortunately quite the opposite.
We can see certain clustered regions of multiple governorates have a collective lower average of grades than others and vice versa.



Geographical heat map of Egypt showing the means of Sciences students performance in their specified subjects for each governorate. The warmer the colour, the higher the mean it has. The map shows that Northern Sinai and Ash Sharqiyah are the top performers while Al Minia and Giza are unfortunately quite the opposite.
We can see certain clustered regions of multiple governorates have a collective lower average of grades than others and vice versa.



Geographical heat map of Egypt showing the means of Linguistic Studies for each governorate. The warmer the colour, the higher the mean it has. The map shows that Northern Sinai, Ad Daqahliah and Ash Sharqiyah are the top performers while Al Minia are unfortunately quite the opposite.
We can see certain clustered regions of multiple governorates have a collective lower average of grades than others and vice versa.

Conclusions

Hypothesis Testing Conclusion:

Since the 3 p-values (p1,p2,p3) are all less than alpha (statistical significance), we can reject the null hypothesis, hence there is a statistically significant difference between the means for the percentages in the governorate of Red Sea and the population.
This is further confirmed by calculating the means from the data set such that:

Percentage Means for All Branches Egypt vs RedSea			
	Literature Branch	Science Branch	Maths Branch
Egypt	58.71	65.75	66.19
RedSea	60.32	63.98	64.24

We can see that there is approximately a 2 average percentage grade difference between RedSea and all of Egypt in each corresponding branch.

General Conclusions:

From what we have extracted from the data above, our conclusions are:

- As expected, many factors go into the grades of high school students that affect them to varying degrees, these include:
    1. Gender: As shown, males typically seem to slightly outperform females, although the 10% proportion difference could be a big factor in this.
    2. Region/Place of residence: Certain regions perform better than others in certain subjects, and some just outright slay the competition like Sinai.
    3. Branch: Maths and Science students exceed the average grades of a Literature student, even in their own specialised subjects like languages.
- From that we know that there are certain points in which support from the government is needed to find out what's going wrong in the under-performing regions and the -on average- lower end of the spectrum literature students. And after identifying these issues, developing a step-by-step concise plan to try and resolve these issues and get these regions and branches up to speed with their peers. This would lead to an overall enhancement of the infrastructure of the educational system.

- More time and a dedicated team of experts could further use this data in the pursuit of determining the issues and devising ways to resolve them.

Potential Issues

- Dealing with a much larger dataset than anything before proved somewhat difficult, especially in the beginning of our work and analysis.
- We likely could have missed some underlying confounding variables that, due to the limitations of this data set, could not be looked deeper into. The difference in regions` grades could have been the byproduct of any of the many possible factors that go into it, could be the services, noise levels, the standard of living, the cleanliness of the water and infinitely many more possible other factors. This would need us to look into more data sets and conduct more deeply thorough research.